#2

PATENT
Docket No. DE9-2000-0096 (270)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of KRIECHBAUM, et al.

Application No. Examiner:

Filed: (Herewith) Group Art Unit:

For: METHOD AND SYSTEM FOR THE AUTOMATIC AMENDMENT OF A SPEECH RECOGNITION VOCABULARIES

CLAIM OF FOREIGN PRIORITY

Box Patent Application
2900 Crystal Drive
Arlington, VA 22202-3513

Sir:

Priority under the International Convention for the Protection of Industrial Property and under 35 U.S.C. §119 is hereby claimed for the above-identified patent application, based upon European Application No. 00127484.4 filed November 29, 2000. A certified copy of European Application No. 00127484.4 is submitted herewith perfecting the Claim of Foreign Priority.

Respectfully submitted,

Date: 11/26/01

Kevin T. Cuenot

Gregory A. Nelson, Registration No. 30,577
Kevin T. Cuenot, Registration No. 46,283
Steven M. Greenberg, Registration No. 44,725
AKERMAN SENTERFITT
222 Lakeview Avenue
Post Office Box 3188
West Palm Beach, FL 33402-3188
Telephone: (561) 653-5000

Docket No. DE9-2000-0096 (270)

Express Mailing Label No. EL 740159505 US

WP068531;1

A2

Europäisches Patentamt

European Patent Office

Office européen des brevets

Bescheinigung

Certificate

Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. **Patent application No.** **Demande de brevet n°**

00127484.4

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

Hatten

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE, 14/05/01
LA HAYE, LE

Europäisches
Patentamt

European
Patent Office

Office européen
des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 00127484.4
Demande n°:

Anmeldetag:
Date of filing: 29/11/00
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):

International Business Machines Corporation
Armonk, NY 10504
UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:

Method and system for the automatic amendment of speech recognition vocabularies

In Anspruch genommene Prioriät(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat:
State:
Pays:

Tag:
Date:
Date:

Aktenzeichen:
File no.
Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:

/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE/TR
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

DESCRIPTION

Method and System for the Automatic Amendment of Speech Recognition Vocabularies

Field of the Invention

The invention generally relates to the field of computer-assisted or computer-based speech recognition and more specifically to a method and system for improving recognition quality of a speech recognition system.

Background of the Invention

Known speech recognition systems (SRSs), in a very simplified view, consist of a database of word pronunciations linked with word spellings. A lot of supplementary mechanisms are used to exploit relevant features of a language and the context of an utterance and thus make the transcription more robust. But all this elaborate mechanisms will not prevent a SRS to fail when either the database of words does not contain a word uttered by a speaker or when a speaker's pronunciation of a word does not agree with the pronunciation entry in the database. Therefore collecting and extending vocabularies is of prime importance for the improvement of SRSs.

Currently vocabularies for SRSs are based on the analysis of large corpora of written documents. For languages where the correspondence between written and spoken language is not bijective, pronunciations have to be entered manually. This is a laborious and costly procedure.

As a first known approach, a mechanism for improving speech recognition through text-based linguistic is disclosed in U.S. Patent 6,064,957. Text data generated from a SRS and a corresponding true transcript of the speech recognition text data are collected and aligned by means of a text aligner. From the differences in alignment a plurality of correction rules is generated by means of a rule generator coupled to the text aligner. The correction rules are then applied by a rule administrator to new text data generated from the SRS. The mechanism does perform only a text-to-text alignment and thus does not take the particular pronunciation of the spoken text into account. It thereupon needs the mentioned rule administrator to apply the rules to new text data and therefore can not be executed fully automatically.

Another approach which allows verbal dictionary updates by end-users of SRSs is known from U.S. Patent 6,078,885. In particular it allows a user to revise the phonetic transcription of words in a phonetic dictionary, or to add transcriptions for words not present in the dictionary. The method determines the phonetic transcription based on the word's spelling and the recorded preferred pronunciation, and updates the dictionary accordingly. Recognition performance is improved through use of the updated dictionary.

The above discussed two approaches thus have in common the drawback that they can not update a speech recognition vocabulary on a large scale of text bodies and with minimum technical and time efforts. Thereupon the known approaches are not fully automated.

Summary of the Invention

It is therefore an object of the present invention to provide a beforehand discussed method and system which enable to improve recognition quality and quantity of a speech recognition system as described beforehand.

It is another object to provide such a method and system which can be executed or performed automatically.

It is another object to provide such a method and system which allow for an improvement of the recognition quality with minimum technical and time efforts.

It is yet another object to provide such a method and system which enable to process large text corpora for updating a speech recognition vocabulary.

The above objects are solved by the features of the independent claims. Advantageous embodiments are subject matter of the subclaims.

The idea underlying the invention is to take an audio realization of a spoken text together with a corresponding allegedly true textual representation (first representation), i.e. an allegedly correct transcription of the audio realization into a text format, to perform a speech recognition on the audio realization thus providing a hypothetic textual representation (second representation) and to look for non-recognized single words using the speech recognition results. These single words then can be used to update a user-dictionary (vocabulary) or pronunciation data obtained by a training of the speech recognition.

It is noted that the true textual representation (true transcript) can be obtained in a digitized format e.g. using known character recognition (OCR) technology.

Further it has been recognized that an automation of the above mentioned mechanism can be achieved by providing a looped procedure where the entire audio realization and both the entire true textual representation and the speech-recognized hypothetic textual representation are aligned to each other so that the true textual representation and the hypothetic textual representation can be aligned to each other likewise. The required information concerning mis-recognized or non-recognized speech segments therefore can be used together with the alignment results in order to locate mis-recognized or non-recognized single words.

It is emphasized that the proposed procedure to identify isolated mis-recognized or non-recognized words in the entire realization and representation and to correlate these words in the audio realization advantageously makes use of an inheritance of the time information from the audio realization and the speech recognized second transcript to the true transcript. Thus the audio signal and both transcriptions can be used to update a word database, a pronunciation database, or both.

The proposed mechanism allows to automate the prementioned vocabulary or dictionary update process and thus to reduce the costs for vocabulary generation e.g. of novel vocabulary domains. The adaptation of a speech recognition system to the idiosyncrasies of a specific speaker is currently an interactive process where the speaker has to correct mis-recognised words. By means of the described invention, such an adaptation too can be automated.

Thereupon the mechanism allows to process large audio or text files and therefore can advantageously used to automatically generate complete vocabularies to start with provided by an average speaker or completely new vocabulary domains to extend an existing vocabulary of a speech recognition system.

Brief Description of the Drawings

In the following the invention will be described in more detail referring to the accompanying drawings from which further features and advantages will become evident.

In the drawings

Fig. 1 is an overview block diagram illustrating a system according to the invention;

Fig. 2 is an overview block diagram of an aligner for aligning a true textual representation and a hypothetical timed transcript in accordance with the invention;

Fig. 3 is an overview block diagram of a classifier that processes the output of the aligner depicted in Fig. 2 in accordance with the invention;

Fig. 4 is an overview block diagram illustrating inheritance of timing information in a system according to the invention;

Fig. 5 is an exemplary data set consisting of a true transcript, a hypothetic transcript provided through speech recognition, and a corresponding timing information output by an aligner in accordance with the invention;

Fig. 6 is an exemplary data set output by a classifier in accordance with the invention;

Fig. 7 shows corresponding data according to a first embodiment of the invention; and

Fig. 8 shows corresponding data according to a second embodiment of the invention.

Detailed Description of the Drawings

Fig. 1 gives an overview of a system and a related procedure according to the invention by way of a block diagram. The procedure starts with a realisation 10, preferably an audio recording of human speech i.e. a spoken text, and a representation 20, preferably a transcription of the spoken text. Many pairs of an audio realisation and a true transcript (resulting from a correct transcription) are publically available e.g. radio featues stored on a storage media like the CD-ROM and their scripts or audio version of text books primarily intended for teaching blind people.

The realisation 10 is first input to a speech recognition engine 50. The textual output of the speech recognition engine 50 and the representation 20 are the aligned by means of an aligner 30 described in more datail hereinbelow referring to Fig. 2. The output of the aligner 30 is passed through a classifier 40 that is described in more detail in Fig. 3. The classifier compares the aligned representation with a transcript produced by a speech recognition engine 50 well-known in the related art and tags all isolated single word recognition errors. An exemplary data set is depicted in Fig. 5.

In case of a first embodiment of the invention, in a next step of the proposed procedure, a selector 60 selects all one word pairs for which the representation and the transcript are different (see also Fig. 6). In case of a second embodiment, word pairs for which the representation and the transcript are similar, are selected for further processing. The selected words together with their corresponding audio signal are then used, in the first embodiment, to update a word database or, in the second embodiment, to update a pronunciation database of a speech recognition system.

Referring to Fig. 2, an aligner used by the present invention in order to align a true representation 100 and a hypothetic timed transcript 110 in a first step expands 120 acronyms and abbreviations, i.e. short forms like 'Mr.' are expanded to the form 'mister' as they are spoken. In a second step all markup is stripped 130 from the texts. For plain ASCII texts this procedure removes all punctuation marks like ';', ',', '.', etc. For texts structered with a markup language all the tags used by the markup language are removed. Special care has to be taken in cases where the transcript has been generated by a SRS system, as in the method and system according to the present invention, working in dictation mode. Hereby the SRS system relies on a command vocabulary to insert punctuation marks. In this case the punctuation has to be expanded to the words used in the command vocabulary, i.e. '.' is replaced by 'full stop'.

After both texts, the time-tagged transcript generated by the

SRS and the representation, have been cleaned as described above, an optiomal word alignment 140 is computed using state-of-the-art techniques as described e.g. in Dan Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge University Press Cambridge 1997. The output of this step is illustrated in Fig. 5 and consists of 4 columns: for each line 600 gives the segments of the representation that align with the segment of the transcript in 610. 620 gives the start time and 630 the end time of the audio signal that gave raise to the transcript 610. It should be noted that due to speech recognition errors the alignment between 610 and 620 is not 1-1 but m-n, i.e. m words of the realisation may be aligned with n words of the transcript.

Fig. 3 is an overview block diagram of the classifer that processes the output of the aligner described above. For all lines 200 in Fig. 5, the classifier adds 210 an additional entry in column 740 as shown in Fig. 6 that specifies whether the correspondence between the representation and the transcript is 1-1 or not. For each line of the aligner output, the classifier tests 220 whether the entry in consists of one word. If this is not true, the value '0' is added 240 in column 740 and the next line of the aligner output is processed. If the entry in column 700 consists only of one word, the same test 230 is applied to the entry in column 710. If this entry too consists only of one word, the value '1' is added 250 in column 740. Otherwise the value '0' is written in 740.

Fig. 4 is an overview block diagram particularly illustrating inheritance of timing information in a system according to the invention. An audio realization, in the present embodiment, is input real-time to a SRS 500 via microphone 510. Alternatively, the audio realization can be provided offline together with a true transcript 520 which is alread be checked for the correctness of the assumed preceding transcription process. It is further assumed that the SRS 500 reveals a timing information for the audio realization. Thus the output of the SRS 500 is a potentially correct transcript 530 that has included the timing information and the timing information 540 itself which can be allegedly accessed separately from the recognized transcript 530.

The original audio realization recorded by the microphone 510 together with the true transcript 520 are input to an aligner 550. A typical output of an aligner 30, 550 is depicted in Fig. 5. It reveals text segments of the true transcript 600 and the recognized transcript 610 together with time stamps representing the start 620 and the stop 630 of each of the text segments. It is emphasized that one part of the text segments like "ich" or "wohl" consists of a single word for both transcripts 600, 610 and the other part of multiple words like "das tue" or "festzuhalten Fuehl".

For the text sample shown in Fig. 5, the corresponding output of a classifier according to the invention is depicted in Fig. 6. The classifier checks the two transcripts 700, 710 (= 600, 610) for text segments that contain identical or similar isolated words and tags 740 those lines accordingly. It is noted hereby that also for similar single words like "Wahn" and "Mann" in both columns 720, 730 the corresponding line is tagged with a "1" bit. The tag information in column 740 now can be used in different ways in accordance with the following two embodiments of the invention.

The first embodiment of the mechanism according to the invention is now described referring to Fig. 7 which enables to automatically update a basic vocabulary of a speech recognition system (SRS). The update, for instance, can be a vocabulary extension of a given domain or supplement of a completely new domain vocabulary to an existing SRS like in the field of medical treatment (radiology etc.). The proposed mechanism selects lines of the output of the classifier (Fig. 7) which comprise a tag bit of "1" but include only not identical single words like "Wahn" and "Mann" in the present example. These single words represent single word recognition errors of the underlying speech recognition engine and therefore can be used in a separate step to update a word database of the underlying SRS.

The second embodiment of the invention allows an automated speaker related adaptation of an existing vocabulary, especially not requiring an active training through the speaker. Hereby only single words where the tag bit equals "1" are selected for which the true transcript (left column) and the recognized transcript (right column) are identical (Fig. 8). These single words represent correctly recognized isolated words and thus can be used in a separate step to update a pronunciation database of an underlying SRS having stored phonetic speaker characteristics.

CLAIMS

1. Method for improving speech recognition, comprising the steps of:

taking a realization and a first representation for the realization;

performing a speech recognition on the realization thus revealing a second representation for the realization;

aligning the first representation and the second representation;

selecting single word pairs for the first representation and the second representation where the first representation and the second representation are different;

updating a word database using the selected single word pairs together with the corresponding aligned realization.

2. Method for improving speech recognition, comprising the steps of:

taking a realization and a first representation for the realization;

performing a speech recognition on the realization thus revealing a second representation for the realization;

aligning the first representation and the second representation;

selecting single word pairs for the first representation and the second representation where the first representation and the second representation are identical;

updating a pronunciation database using the selected single word pairs together with the corresponding aligned realization.

3. Method according to claim 1 or 2, wherein the step of selecting single word pairs uses resulting speech recognition information.

4. Method according to claim 2 or 3, wherein the step of updating the pronunciation database comprises the further step of comparing the recognition quality of the speech recognition on the realization with the recognition quality of a corresponding single word entry existing in the pronunciation database.

5. Method according to any of the preceding claims, wherein the step of comparing the aligned first representation with the second representation comprises the further step of tagging all segments of the first and the second representation were both the first and the second representation consist of a single word.

6. Method according to any of the preceding claims, wherein the step of aligning the first representation and the second representation reveals time information concerning the alignment between the realization and the first representation.

7. A system for improving speech recognition of a speech recognizer transcribing a realization into a representation, the system comprising:

an aligner for aligning a first representation and a second representation revealed from the speech recognizer;

a classifier for comparing the aligned first representation with the aligned second representation;

a selector for selecting single word pairs where the aligned first representation and the aligned second representation are different or identical.

8. System according to claim 7, where the aligner further comprises means for generating time information concerning time alignment between the first representation and the second representation.

9. System according to claim 7 or 8, where the classifier further comprises means for tagging all segments of the first representation and the second representation were both the first representation and the second representation consist of a single word.

10. System according to any of claims 7 to 9, further comprising means for updating a word database and/or a pronunciation database using the selected single word pairs.

THIS PAGE BLANK (USPTO)

ABSTRACT

Disclosed is a mechanism to improve speech recognition which uses an existing audio realization of a spoken text and a true textual representation of the spoken text. The audio realization and the true textual representation are aligned in order to reveal time stamps. Further a speech recognition is performed on the audio realization thus revealing a hypothetic textual representation for the audio realization. Afterwards the aligned true textual representation is compared with the hypothetic textual representation revealed by the speech recognition and single word pairs for the true and the hypothetic textual representation are selected where the true and the hypothetic textual representation, according to a first embodiment, are different. According to a second embodiment, single word pairs are selected where the true and the hypothetic textual representation are identical. Finally a word database or, according to the second embodiment, a pronunciation database is updated using the selected single word pairs together with the corresponding aligned audio realization.
(Fig. 1)

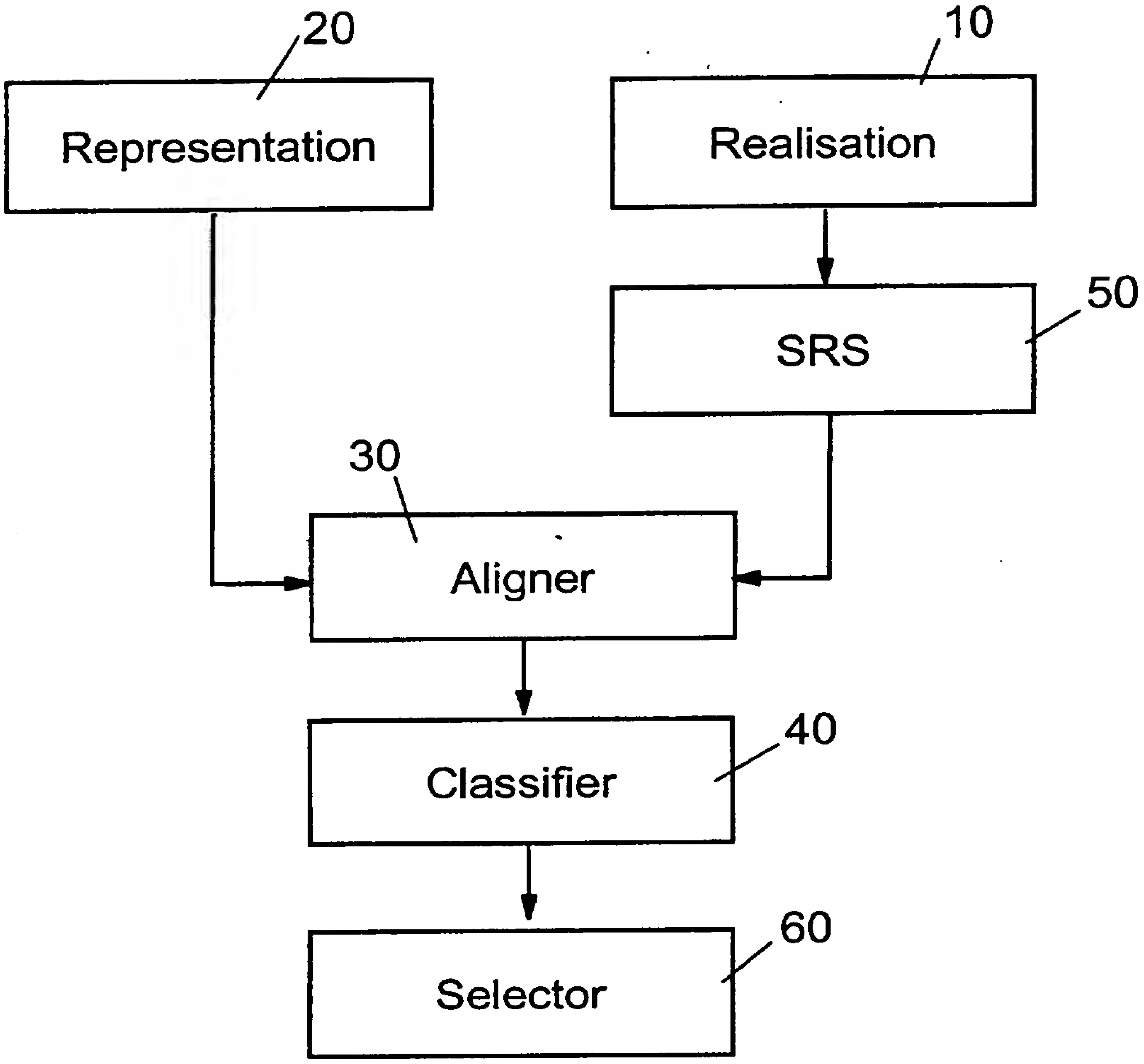
20
Representation
10
Realisation
50
SRS
30
Aligner
40
Classifier
60
Selector

FIG. 1

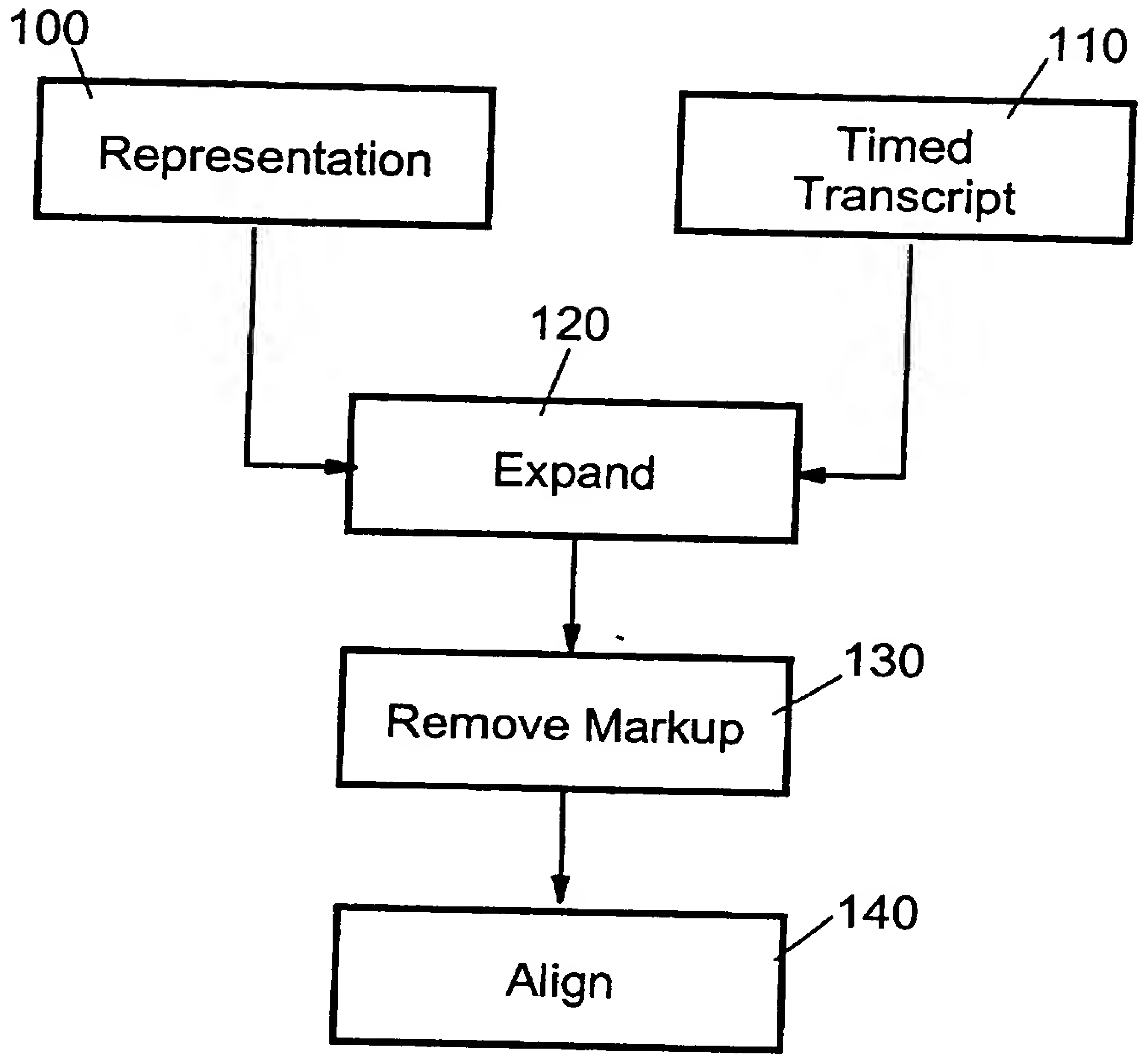
100
Representation
110
Timed Transcript
120
Expand
130
Remove Markup
140
Align

FIG.2

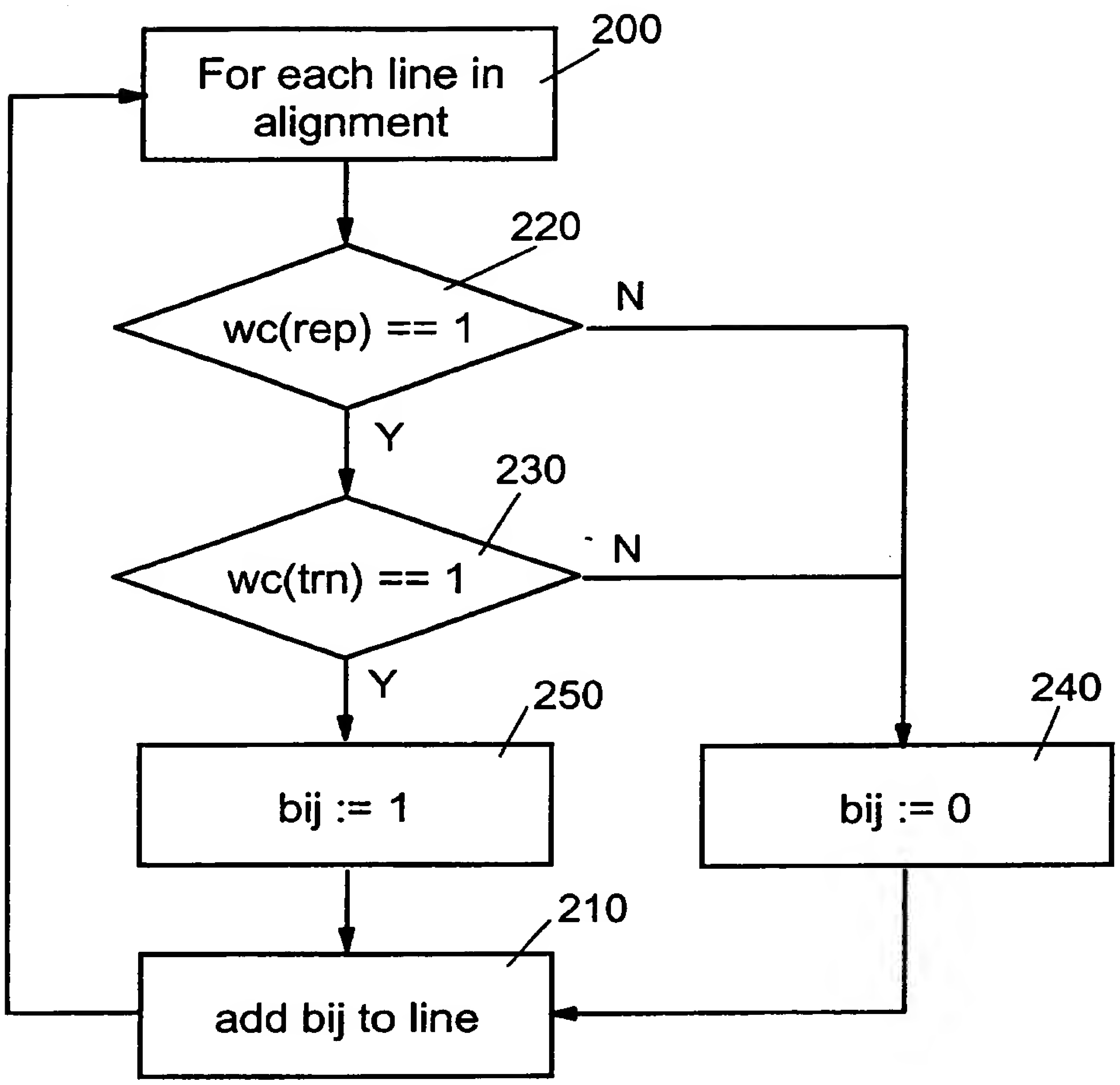
For each line in alignment
200
220
wc(rep) == 1
N
Y
230
wc(trn) == 1
N
Y
250
240
bij := 1
bij := 0
210
add bij to line

FIG. 3

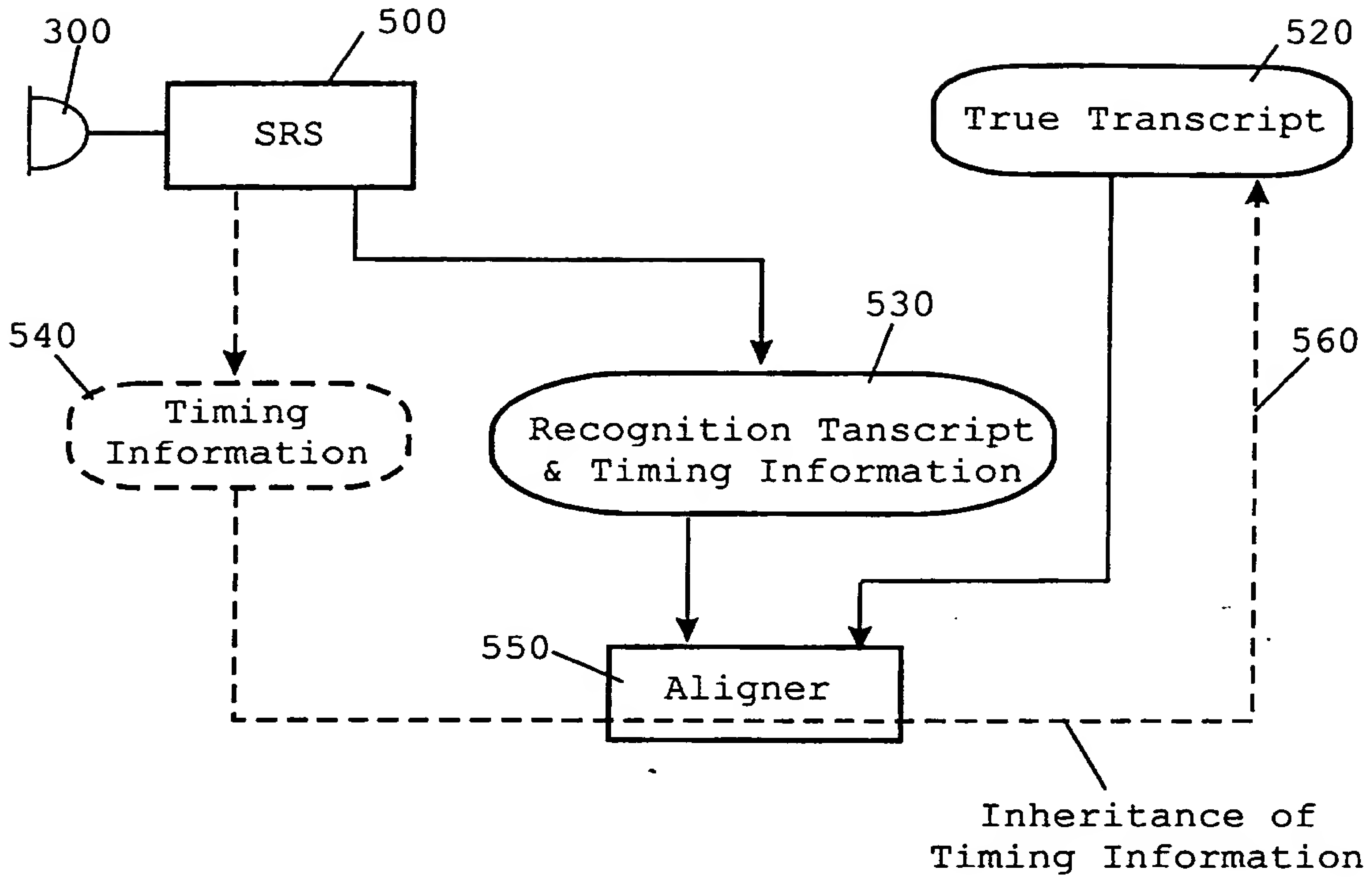
300
500
SRS
520
True Transcript
530
540
560
Timing Information
Recognition Tanscript & Timing Information
550
Aligner
Inheritance of Timing Information

FIG. 4

600	610	620	630
Representation	Transcript	Start	Stop
Versuch	das tue	1137	1436
ich	ich	1436	1666
wohl	wohl	1666	1925
euch	euch	1925	2244
diesmal	diesmal	2244	2789
festzuhalten Fuehl	ist es zu alt zum Kueche	2789	6305
ich	ich	6305	6375
mein Herz	merkst	6375	7004
noch	noch	7004	7473
jenem	jenem	7473	8281
Wahn	Mann	8281	8730
geneigt	geneigt	8730	9258

FIG. 5

700	710	720	730	740
Representation	Transcript	Start	Stop	1-1
Versuch	das tue	1137	1436	0
ich	ich	1436	1666	1
wohl	wohl	1666	1925	1
euch	euch	1925	2244	1
diesmal	diesmal	2244	2789	1
festzuhalten Fuehl	ist es zu alt zum Kueche	2789	6305	0
ich	ich	6305	6375	1
mein Herz	merkst	6375	7004	0
noch	noch	7004	7473	1
jenem	jenem	7473	8281	1
Wahn	Mann	8281	8730	1
geneigt	geneigt	8730	9258	1

FIG. 6

Representation	Transcript	Start	Stop	1-1
Wahn	Mann	8281	8730	1

FIG. 7

Representation	Transcript	Start	Stop	1-1
ich	ich	1436	1666	1
wohl	wohl	1666	1925	1
euch	euch	1925	2244	1
diesmal	diesmal	2244	2789	1
ich	ich	6305	6375	1
noch	noch	7004	7473	1
jenem	jenem	7473	8281	1
geneigt	geneigt	8730	9258	1

FIG. 8